Published online 12 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D33–D37***
doi:10.1093/nar/gkr1006

The International Nucleotide Sequence Database Collaboration

Ilene Karsch-Mizrachi[1,*], Yasukazu Nakamura[2] and Guy Cochrane[3], on behalf of the International Nucleotide Sequence Database Collaboration

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD 20892, USA, [2]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima 411-8510, Japan and [3]EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received October 18, 2011; Accepted October 19, 2011

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 8, 2015

ABSTRACT

The members of the International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org) set out to capture, preserve and present globally comprehensive public domain nucleotide sequence information. The work of the long-standing collaboration includes the provision of data formats, annotation conventions and routine global data exchange. Among the many developments to INSDC resources in 2011 are the newly launched BioProject database and improved handling of assembly information. In this article, we outline INSDC services and update the reader on developments in 2011.

INTRODUCTION

With the arrival of 2012, the International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org) enters its third decade of existence. This fundamental global data exchange collaboration serves as a model for data sharing across and beyond the life sciences. Together, the three INSDC partners, the DNA Databank of Japan (DDBJ; http://www.ddbj.nig.ac.jp/) at the National Institute for Genetics in Mishima, Japan; the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI; http://www.ebi.ac.uk/ena) in Hinxton, UK; and the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/genbank/) in Bethesda, Maryland, USA, capture, preserve and present the permanent scientific record for nucleic acid sequencing and associated information. Over the years of collaboration, the scope of sequencing activity has grown enormously and INSDC has developed a mandate that covers the richness of the domain, including repositories and services for raw data (the global Trace Archives and Sequence Read Archive [SRA, (1)], assembly data, experimental design details, taxonomic information, functional annotation and large sequencing project information.

The smooth operation of the INSDC depends on the existence and maintenance of a number of collaborative instruments and policies that are documented on the INSDC website. These include the Feature Table Definitions and Status Definition documents, the global accessioning system, rapid data exchange between partners and core INSDC policies, such as support for mandatory data submission in support of literature publication, developed in collaboration with the INSDC International Advisory Committee. Further details of collaborative instruments and policies have been presented (2).

In this article, we outline the recent developments and report in detail on two major areas of development, the launch of the BioProject database and improvements to handling of assembly information, that allow INSDC partners to respond to changing user needs. Finally, we provide an update on INSDC content in 2011.

RECENT DEVELOPMENTS

Members of the INSDC meet annually to discuss issues related to maintaining the sequence archives. These issues range from the addition of feature or qualifier elements to the feature tables present in the Flat File report format in the traditional archive records to policy issues and strategies for dealing with the increasing sequence data to be archived. In 2011, the annual meeting was held in Osaka, Japan. At the meeting, we finalized a number of important changes to the document describing the availability of sequences across INSDC partner sites (http://www.insdc.org/documents/status_document.html).

*To whom correspondence should be addressed. Tel: +301 435 5929; Fax: +301 480 2918; Email: mizrachi@ncbi.nlm.nih.gov

Published by Oxford University Press 2011.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

This document describes such concepts as fully public data, data held confidential prior to publication and data replaced as updated improved data become available. It also defines instances where sequences that were once public may be removed.

A number of new feature keys and qualifiers were adopted. Beyond those that relate to changes to the way in which assembly information is handled (see below), the new feature keys centromere and telomere were adopted to support specific labeling of these subchromosomal regions. In addition, changes in usage of /pseudo and the introduction of /pseudogene were agreed. Details of these updates are announced at INSDC partner sites and are laid out, including dates and timelines for implementation, in the October Feature Table Definitions document (http://www.insdc.org/documents/feature_table.html).

Discussion of SRA centered on schema simplifications, data compression and preparation for sustainable mechanisms of data exchange. Details have been published in this issue (1).

BIOPROJECT DATABASE

BioProject was released as a collaborative resource in 2011. It is becoming increasingly important to collect metadata associated with large-scale sequencing initiatives

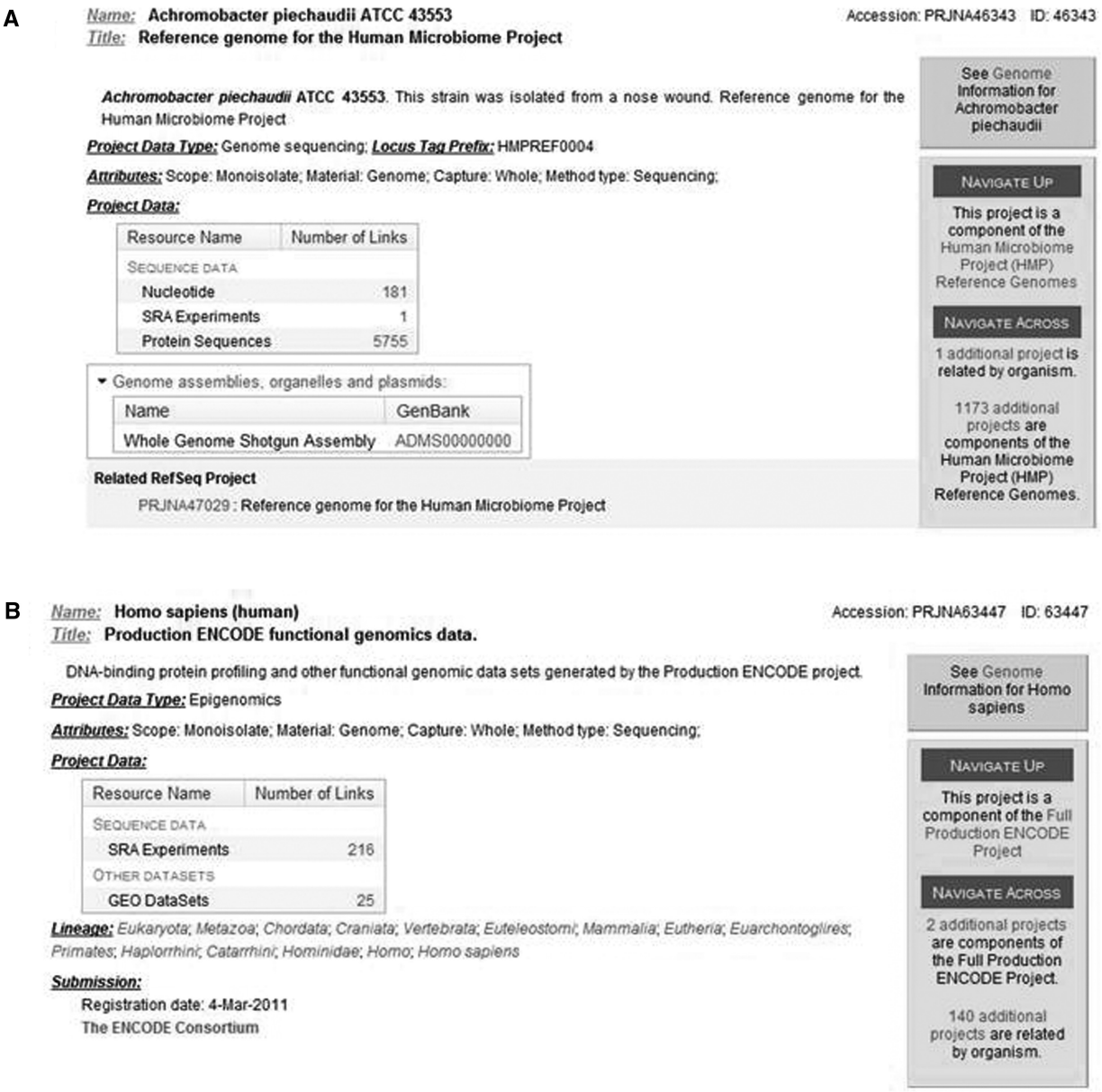

Figure 1. Sample BioProject records, shown from the NCBI website, for (**A**) a genome sequencing project and (**B**) an epigenomic project. Linkages are created from the BioProject record to the resources containing data and from the BioProject record to other BioProject records that are part of the same initiative or contain related data.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 8, 2015

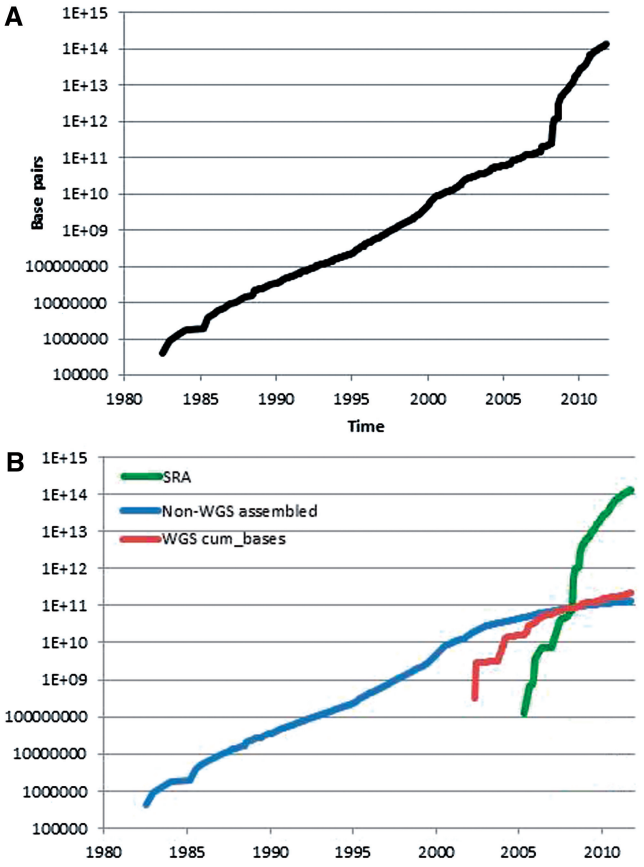

Figure 2. (**A**) Cumulative base pairs in INSDC over time, excluding the Trace Archive (raw data from capillary sequencing platforms). (**B**) Base pairs in INSDC over time since 1980, broken down into selected data components. Cumulative data volume in base pairs broken down into assembled sequence (whole genome shotgun methods and others) and raw next-generation-sequence data.

and for the archives to aggregate this information to allow users to easily access the information. For the past few years, genome sequence submissions to INSDC contained a Genome Project identifier with links to the Genome Project database. The Genome Project record held information about genomic molecules for the sequencing effort of a particular species with links to the INSDC accessions. It became increasingly clear that accessions from other types of initiatives needed to be aggregated and cited in a central database. This is especially true when there are disparate types of data in a single funded initiative. The new BioProject database (3), which has superseded Genome Project, enables submitters to describe initiatives that include data from transcriptome, metagenome, targeted locus and multi-isolate studies that initiate from a single organization, consortium or funding initiative. Additionally, linkages can be created between different project types to provide users with a single entry point to the data. Linkages may also be established between initiatives with data deposited in the INSDC sequence archives and repositories that house analysis of raw sequence output, such as GEO (4), ArrayExpress (5) and dbGaP. Sample BioProject records with links to different data types are shown in Figure 1.

HANDLING OF ASSEMBLY INFORMATION

In 2008, an agreement was reached by the major genome browser and annotation groups, Ensembl, NCBI and UCSC, whereby each site would be required to consistently display the same assemblies with the same coordinate systems (http://www.ensembl.org/info/about/legal/browser_agreement.html). The group agreed that: (i) data displayed in the assembly should be displayed only after it has been released in INSDC. (ii) The sequence identifiers used in the browser and publicly distributed via FTP should be correlated with the INSDC records. (iii) All browsers will refer to any given assembly by the same name, preferably a submitter approved name. Prior to this agreement, in many instances, the browsers were annotating and displaying different versions of genome assemblies which made it difficult to compare data originating from different data providers. In light of this, INSDC must work with the genome centers producing the data to ensure genomes are submitted and released in a timely manner so that they may be available for use by the browsers.

In the genomics community, the format of submission and exchange for genome assembly data is the AGP (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/agp/AGP_Specification.shtml). This tabular format describes the assembly of a genomic object, i.e. scaffold and/or chromosome, from component sequence records. Each row of the AGP describes a piece of the object and information about the coordinates, type of component or gap and linkages are specified in the columns. At INSDC, the constructions delineated in the AGP file are built as CON Flat File records, which contain instructions for how to build a scaffolded sequence from the component sequence records.

The AGP specification captures richer information about the type of gap, but this has to-date been discarded in the INSDC Flat File view of the CON record. Additionally, the AGP specification will be extended to explicitly indicate the evidence for linkage between two components. This information would also be potentially lost in the CON representation of the sequence assembly. Therefore, to better capture this information, INSDC will improve the format and display of the CON record by introducing a new feature key, assembly_gap and two new qualifiers, /gap_type and /linkage_evidence to capture the additional information that is included in AGP files that was previously lost from the INSDC records. This change will also enable genome centers and the browser groups to interconvert losslessly the INSDC CON representation and the AGP files.

CONTENT IN 2011

In 2011, INSDC databases have grown overall around 2-fold in terms of the number of bases (Figure 2a). Behind this absolute growth are increases in the numbers of assembled sequences of 14% (Figure 3). The rate of accumulation of assembled sequences is decreasing over time as the rate of accumulation of SRA bases increases. This is evident especially for bulk data sequence

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 8, 2015

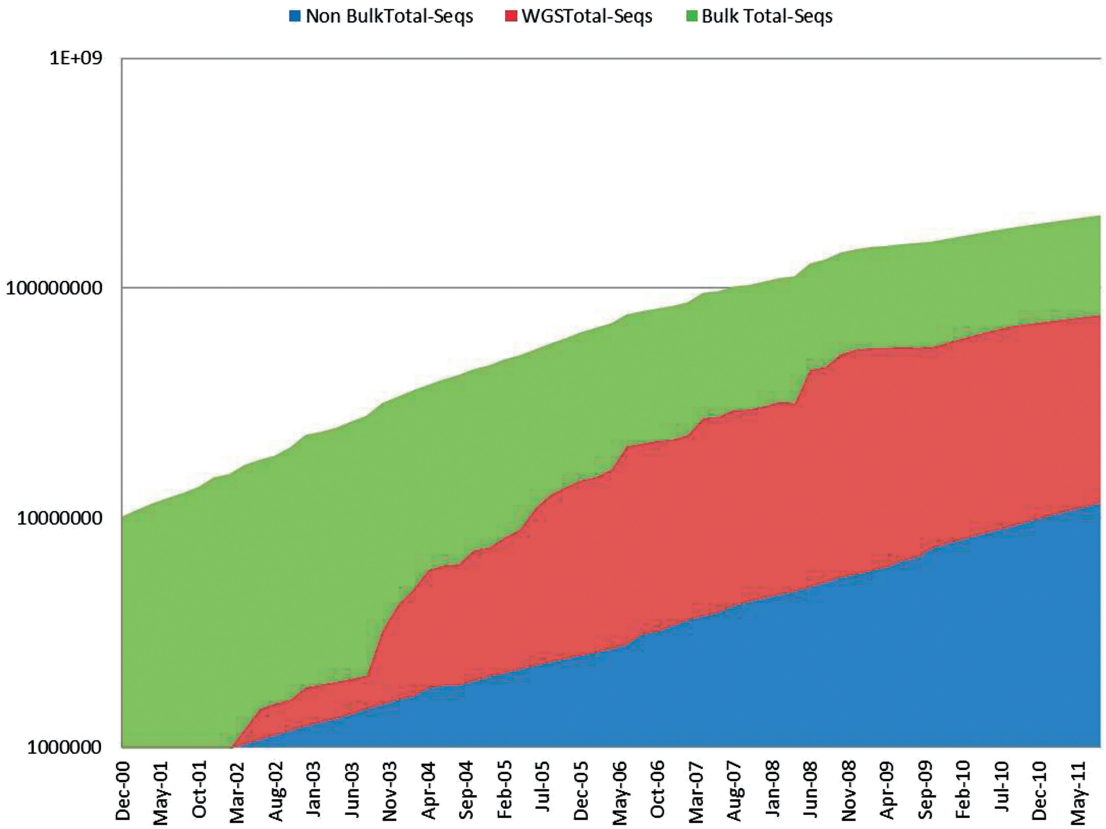

Figure 3. Cumulative growth in the number of sequences included in the traditional INSDC sequence archives over time. Bulk sequence data includes non-WGS bulk submission types i.e. EST, GSS, Patent and Transcriptome Shotgun Assembly (TSA). WGS includes the number of sequence overlap contigs. Non-bulk data is the remainder.

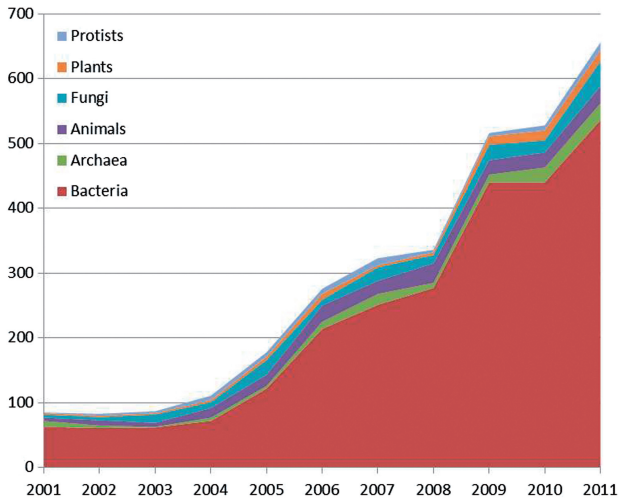

Figure 4. Growth in genomes. The layered chart shows the number of new species with genomes entered into INSDC databases over time by taxonomic group. The 2011 time point includes data released in the first 9 months.

data submissions, such as Expressed Sequence Tags (EST) and Genome Survey Sequences (GSS) where the SRA deposition is more appropriate. To a lesser extent, Whole Genome Shotgun (WGS) is also being deposited at a slower rate than previously even though the contribution of WGS sequences is still a significant (Figures 2b and 3) component of traditional sequences.

Despite this slowing of growth in assembled sequence submissions to INSDC databases, it is clear that the catalogue of public domain genomes continues to grow rapidly and is expected to continue to grow as the availability and quality of next-generation sequence is enhanced. In the first 9 months of 2011, over 1200 new genomes were deposited, more than half of these were genomes from species that were not previously sequenced. Figure 4 shows the growth in the number of new species with genomes deposited in INSDC. We anticipate at least a 10-fold increase in the number of genomes submitted to INSDC in the next 2 years.

FUNDING

NCBI by the Intramural Research Program of the National Institutes of Health; National Library of Medicine; European Nucleotide Archive by the European Molecular Biology Laboratory, the Wellcome Trust, the FP7 Programme of the European Commission and the Biotechnology and Biological Sciences Research Council; at DDBJ by the Ministry of Education, Culture, Sports, Science and Technology of Japan. Funding for

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 8, 2015

open access charge: the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
2. Cochrane,G., Karsch-Mizrachi,I. and Nakamura,Y. (2011) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **39**, D15–D18.
3. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
4. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
5. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,I.E., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 8, 2015